DECLARATION OF PAUL POLAKIS, Ph.D.

I, Paul Polakis, Ph.D., declare and say as follows:

1. I was awarded a Ph.D. by the Department of Biochemistry of the Michigan State University in 1984. My scientific Curriculum Vitae is attached to and forms part of this Declaration (Exhibit A).

2. I am currently employed by Genentech, Inc. where my job title is Staff Scientist. Since joining Genentech in 1999, one of my primary responsibilities has been leading Genentech's Tumor Antigen Project, which is a large research project with a primary focus on identifying tumor cell markers that find use as targets for both the diagnosis and treatment of cancer in humans.

3. As part of the Tumor Antigen Project, my laboratory has been analyzing differential expression of various genes in tumor cells relative to normal cells. The purpose of this research is to identify proteins that are abundantly expressed on certain tumor cells and that are either (i) not expressed, or (ii) expressed at lower levels, on corresponding normal cells. We call such differentially expressed proteins "tumor antigen proteins". When such a tumor antigen protein is identified, one can produce an antibody that recognizes and binds to that protein. Such an antibody finds use in the diagnosis of human cancer and may ultimately serve as an effective therapeutic in the treatment of human cancer.

4. In the course of the research conducted by Genentech's Tumor Antigen Project, we have employed a variety of scientific techniques for detecting and studying differential gene expression in human tumor cells relative to normal cells, at genomic DNA, mRNA and protein levels. An important example of one such technique is the well known and widely used technique of microarray analysis which has proven to be extremely useful for the identification of mRNA molecules that are differentially expressed in one tissue or cell type relative to another. In the course of our research using microarray analysis, we have identified approximately 200 gene transcripts that are present in human tumor cells at significantly higher levels than in corresponding normal human cells. To date, we have generated antibodies that bind to about 30 of the tumor antigen proteins expressed from these differentially expressed gene transcripts and have used these antibodies to quantitatively determine the level of production of these tumor antigen proteins in both human cancer cells and corresponding normal cells. We have then compared the levels of mRNA and protein in both the tumor and normal cells analyzed.

5. From the mRNA and protein expression analyses described in paragraph 4 above, we have observed that there is a strong correlation between changes in the level of mRNA present in any particular cell type and the level of protein

expressed from that mRNA in that cell type. In approximately 80% of our observations we have found that increases in the level of a particular mRNA correlates with changes in the level of protein expressed from that mRNA when human tumor cells are compared with their corresponding normal cells.

6. Based upon my own experience accumulated in more than 20 years of research, including the data discussed in paragraphs 4 and 5 above and my knowledge of the relevant scientific literature, it is my considered scientific opinion that for human genes, an increased level of mRNA in a tumor cell relative to a normal cell typically correlates to a similar increase in abundance of the encoded protein in the tumor cell relative to the normal cell. In fact, it remains a central dogma in molecular biology that increased mRNA levels are predictive of corresponding increased levels of the encoded protein. While there have been published reports of genes for which such a correlation does not exist, it is my opinion that such reports are exceptions to the commonly understood general rule that increased mRNA levels are predictive of corresponding increased levels of the encoded protein.

7. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information or belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful statements may jeopardize the validity of the application or any patent issued thereon.

Dated: 5/07/04

By: Paul Polakis

Paul Polakis, Ph.D.

SV 2031808 v1

CURRICULUM VITAE

PAUL G. POLAKIS
Staff Scientist
Genentech, Inc
1 DNA Way, MS#40
S. San Francisco, CA 94080

EDUCATION:

Ph.D., Biochemistry, Department of Biochemistry,
Michigan State University (1984)

B.S., Biology. College of Natural Science, Michigan State University (1977)

PROFESSIONAL EXPERIENCE:

2002-present	Staff Scientist, Genentech, Inc S. San Francisco, CA
1999- 2002	Senior Scientist, Genentech, Inc., S. San Francisco, CA
1997 -1999	Research Director Onyx Pharmaceuticals, Richmond, CA
1992- 1996	Senior Scientist, Project Leader, Onyx Pharmaceuticals, Richmond, CA
1991-1992	Senior Scientist, Chiron Corporation, Emeryville, CA.
1989-1991	Scientist, Cetus Corporation, Emeryville CA.
1987-1989	Postdoctoral Research Associate, Genentech, Inc., South SanFrancisco, CA.
1985-1987	Postdoctoral Research Associate, Department of Medicine, Duke University Medical Center, Durham, NC

1984-1985	Assistant Professor, Department of Chemistry, Oberlin College, Oberlin, Ohio
1980-1984	Graduate Research Assistant, Department of Biochemistry, Michigan State University East Lansing, Michigan

PUBLICATIONS:

1. Polakis, P G. and Wilson, J. E. 1982 Purification of a Highly Bindable Rat Brain Hexokinase by High Performance Liquid Chromatography. **Biochem. Biophys. Res. Commun**. 107, 937-943.

2. Polakis, P.G. and Wilson, J. E. 1984 Proteolytic Dissection of Rat Brain Hexokinase: Determination of the Cleavage Pattern during Limited Digestion with Trypsin. **Arch. Biochem. Biophys**. 234, 341-352.

3. Polakis, P. G. and Wilson, J. E. 1985 An Intact Hydrophobic N-Terminal Sequence is Required for the Binding Rat Brain Hexokinase to Mitochondria. **Arch. Biochem. Biophys**. 236, 328-337.

4. Uhing, R.J., **Polakis,P.G**. and Snyderman, R. 1987 Isolaton of GTP-binding Proteins from Myeloid HL60 Cells. **J. Biol. Chem**. 262, 15575-15579.

5. Polakis, P.G., Uhing, R.J. and Snyderman, R. 1988 The Formylpeptide Chemoattractant Receptor Copurifies with a GTP-binding Protein Containing a Distinct 40 kDa Pertussis Toxin Substrate. **J. Biol. Chem**. 263, 4969-4979.

6. Uhing, R. J., Dillon, S., **Polakis, P. G.**, Truett, A. P. and Snyderman, R. 1988 Chemoattractant Receptors and Signal Transduction Processes in Cellular and Molecular Aspects of Inflammation (Poste, G. and Crooke, S. T. eds.) pp 335-379.

7. Polakis, P.G., Evans, T. and Snyderman 1989 Multiple Chromatographic Forms of the Formylpeptide Chemoattractant Receptor and their Relationship to GTP-binding Proteins. **Biochem. Biophys. Res. Commun**. 161, 276-283.

8. Polakis, P. G., Snyderman, R. and Evans, T. 1989 Characterization of G25K, a GTP-binding Protein Containing a Novel Putative Nucleotide Binding Domain. **Biochem. Biophys. Res. Comun.** 160, 25-32.

9. Polakis, P., Weber,R.F., Nevins,B., Didsbury, J. Evans,T. and Snyderman, R. 1989 Identification of the ral and rac1 Gene Products, Low Molecular Mass GTP-binding Proteins from Human Platelets. **J. Biol. Chem**. 264, 16383-16389.

10. Snyderman, R., Perianin, A., Evans, T., **Polakis, P.** and Didsbury, J. 1989 G Proteins and Neutrophil Function. In ADP-Ribosylating Toxins and G Proteins:

Insights into Signal Transduction. (J. Moss and M. Vaughn, eds.) Amer. Soc. Microbiol. pp. 295-323.

11. Hart, M.J., **Polakis, P.G.**, Evans, T. and Cerrione, R.A. 1990 The Identification and Charaterization of an Epidermal Growth Factor-Stimulated Phosphorylation of a Specific Low Molecular Mass GTP-binding Protein in a Reconstituted Phospholipid Vesicle System. **J. Biol. Chem**. 265, 5990-6001.

12. Yatani, A., Okabe, K., **Polakis, P.** Halenbeck, R. McCormick, F. and Brown, A. M. 1990 ras p21 and GAP Inhibit Coupling of Muscarinic Receptors to Atrial K^+ Channels. **Cell**. 61, 769-776.

13. Munemitsu, S., Innis, M.A., Clark, R., McCormick, F., Ullrich, A. and **Polakis, P.G.** 1990 Molecular Cloning and Expression of a G25K cDNA, the Human Homolog of the Yeast Cell Cycle Gene CDC42. **Mol. Cell. Biol**. 10, 5977-5982.

14. Polakis, P.G. Rubinfeld, B. Evans, T. and McCormick, F. 1991 Purification of Plasma Membrane-Associated GTPase Activating Protein (GAP) Specific for rap-1/krev-1 from HL60 Cells. **Proc. Natl. Acad. Sci. USA** 88, 239-243.

15. Moran, M. F., **Polakis, P.**, McCormick, F., Pawson, T. and Ellis, C. 1991 Protein Tyrosine Kinases Regulate the Phosphorylation, Protein Interactions, Subcellular Distribution, and Activity of p21ras GTPase Activating Protein. **Mol. Cell. Biol**. 11, 1804-1812

16. Rubinfeld, B., Wong, G., Bekesi, E. Wood, A. McCormick, F. and **Polakis, P. G.** 1991 A Synthetic Peptide Corresponding to a Sequence in the GTPase Activating Protein Inhibits $p21^{ras}$ Stimulation and Promotes Guanine Nucleotide Exchange. **Internatl. J. Peptide and Prot. Res**. 38, 47-53.

17. Rubinfeld, B., Munemitsu, S., Clark, R., Conroy, L., Watt, K., Crosier, W., McCormick, F., and **Polakis, P.** 1991 Molecular Cloning of a GTPase Activating Protein Specific for the Krev-1 Protein $p21^{rap1}$. **Cell** 65, 1033-1042.

18. Zhang, K. Papageorge, A., G., Martin, P., Vass, W. C., Olah, Z., **Polakis, P.**, McCormick, F. and Lowy, D, R. 1991 Heterogenous Amino Acids in RAS and Rap1A Specifying Sensitivity to GAP Proteins. **Science** 254, 1630-1634.

19. Martin, G., Yatani, A., Clark, R., **Polakis, P.**, Brown, A. M. and McCormick, F. 1992 GAP Domains Responsible for $p21^{ras}$-dependent Inhibition of Muscarinic Atrial K^+ Channel Currents. **Science** 255, 192-194.

20. McCormick, F., Martin, G. A., Clark, R., Bollag, G. and **Polakis, P** . 1992 Regulation of p21ras by GTPase Activating Proteins. Cold Spring Harbor **Symposia on Quantitative Biology**. Vol. 56, 237-241.

21. Pronk, G. B., **Polakis, P.**, Wong, G., deVries-Smits, A. M., Bos J. L. and McCormick, F. 1992 $p60^{v\text{-}src}$ Can Associate with and Phosphorylate the $p21^{ras}$ GTPase Activating Protein. **Oncogene** 7,389-394.

22. Polakis P. and McCormick, F. 1992 Interactions Between $p21^{ras}$ Proteins and Their GTPase Activating Proteins. In **Cancer Surveys** (Franks, L. M., ed.) 12, 25-42.

23. Wong, G., Muller, O., Clark, R., Conroy, L., Moran, M., **Polakis,P.** and McCormick, F. 1992 Molecular coloning and nucleic acid binding properties of the GAP-associated tyrosine phosphoprotein p62. **Cell** 69, 551-558.

24. Polakis, P., Rubinfeld, B. and McCormick, F. 1992 Phosphorylation of rap1GAP in vivo and by cAMP-dependent Kinase and the Cell Cycle $p34^{cdc2}$ Kinase in vitro. **J. Biol. Chem**. 267, 10780-10785.

25. McCabe, P.C., Haubrauck, H., **Polakis, P.**, McCormick, F., and Innis, M. A. 1992 Functional Interactions Between $p21^{rap1A}$ and Components of the Budding pathway of *Saccharomyces cerevisiae*. **Mol. Cell. Biol**. 12, 4084-4092.

26. Rubinfeld, B., Crosier, W.J., Albert, I., Conroy,L., Clark, R., McCormick, F. and **Polakis, P**. 1992 Localization of the rap1GAP Catalytic Domain and Sites of Phosphorylation by Mutational Analysis. **Mol. Cell . Biol**. 12, 4634-4642.

27. Ando, S., Kaibuchi, K., Sasaki, K., Hiraoka, T., Nishiyama, T., Mizuno, T., Asada, M., Nunoi, H., Matsuda, I., Matsuura, Y., **Polakis, P.**, McCormick, F. and Takai, Y. 1992 Post-translational processing of rac p21s is important both for their interaction with the GDP/GTP exchange proteins and for their activation of NADPH oxidase. **J. Biol. Chem**. 267, 25709-25713.

28. Janoueix-Lerosey, I., **Polakis, P.**, Tavitian, A. and deGunzberg, J. 1992 Regulation of the GTPase activity of the ras-related rap2 protein. **Biochem. Biophys. Res. Commun**. 189, 455-464.

29. Polakis, P. 1993 GAPs Specific for the rap1/Krev-1 Protein. in GTP-binding Proteins: the *ras*-superfamily. (J.C. LaCale and F. McCormick, eds.) 445-452.

30. Polakis, P. and McCormick, F. 1993 Structural requirements for the interaction of $p21^{ras}$ with GAP, exchange factors, and its bological effector target. **J. Biol Chem**. 268, 9157-9160.

31. Rubinfeld, B., Souza, B. Albert, I., Muller, O., Chamberlain, S., Masiarz, F., Munemitsu, S. and **Polakis, P.** 1993 Association of the APC gene product with beta- catenin. **Science** 262, 1731-1734.

32. Weiss, J., Rubinfeld, B., **Polakis, P.**, McCormick, F. Cavenee, W. A. and Arden, K. 1993 The gene for human rap1-GTPase activating protein (rap1GAP) maps to chromosome 1p35-1p36.1. **Cytogenet. Cell Genet**. 66, 18-21.

33. Sato, K. Y., **Polakis, P.**, Haubruck, H., Fasching, C. L., McCormick, F. and Stanbridge, E. J. 1994 Analysis of the tumor suppressor acitvity of the K-rev gene in human tumor cell lines. **Cancer Res**. 54, 552-559.

34. Janoueix-Lerosey, I., Fontenay, M., Tobelem, G., Tavitian, A., **Polakis, P.** and DeGunzburg, J. 1994 Phosphorylation of rap1GAP during the cell cycle. **Biochem. Biophys. Res. Commun**. 202, 967-975

35. Munemitsu, S., Souza, B., Mueller, O., Albert, I., Rubinfeld, B., and **Polakis, P.** 1994 The APC gene product associates with microtubules in vivo and affects their assembly in vitro. **Cancer Res**. 54, 3676-3681.

36. Rubinfeld, B. and **Polakis, P.** 1995 Purification of baculovirus produced rap1GAP. **Methods Enz**. 255,31

37. Polakis, P. 1995 Mutations in the APC gene and their implications for protein structure and function. **Current Opinions in Genetics and Development** 5, 66-71

38. Rubinfeld, B., Souza, B., Albert, I., Munemitsu, S. and **Polakis P.** 1995 The APC protein and E-cadherin form similar but independent complexes with α-catenin, β-catenin and Plakoglobin. **J. Biol. Chem**. 270, 5549-5555

39. Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B., and **Polakis, P.** 1995 Regulation of intracellular β-catenin levels by the APC tumor suppressor gene. **Proc. Natl. Acad. Sci**. 92, 3046-3050.

40. Lock, P., Fumagalli, S., **Polakis, P.** McCormick, F. and Courtneidge, S. A. 1996 The human p62 cDNA encodes Sam68 and not the rasGAP-associated p62 protein. **Cell** 84, 23-24.

41. Papkoff, J., Rubinfeld, B., Schryver, B. and **Polakis, P.** 1996 Wnt-1 regulates free pools of catenins and stabilizes APC-catenin complexes. **Mol. Cell. Biol**. 16, 2128-2134.

42. Rubinfeld, B., Albert, I., Porfiri, E., Fiol, C., Munemitsu, S. and **Polakis, P.** 1996 Binding of GSK3β to the APC-β-catenin complex and regulation of complex assembly. **Science** 272, 1023-1026.

43. Munemitsu, S., Albert, I., Rubinfeld, B. and **Polakis, P.** 1996 Deletion of amino-terminal structure stabilizes β-catenin in vivo and promotes the hyperphosphorylation of the APC tumor suppressor protein. **Mol. Cell. Biol**.16, 4088-4094.

44. Hart, M. J., Callow, M. G., Sousa, B. and **Polakis P.** 1996 IQGAP1, a calmodulin binding protein witha rasGAP related domain, is a potential effector for cdc42Hs. **EMBO J.** 15, 2997-3005.

45. Nathke, I. S., Adams, C. L., **Polakis, P.**, Sellin, J. and Nelson, W. J. 1996 The adenomatous polyposis coli (APC) tumor suppressor protein is localized to plasma membrane sites involved in active epithelial cell migration. **J. Cell. Biol.** 134, 165-180.

46. Hart, M. J., Sharma, S., elMasry, N., Qui, R-G., McCabe, P., **Polakis, P.** and Bollag, G. 1996 Identification of a novel guanine nucleotide exchange factor for the rho GTPase. **J. Biol. Chem.** 271, 25452.

47. Thomas JE, Smith M, Rubinfeld B, Gutowski M, Beckmann RP, and **Polakis P.** 1996 Subcellular localization and analysis of apparent 180-kDa and 220-kDa proteins of the breast cancer susceptibility gene, BRCA1. **J. Biol. Chem.** 1996 271, 28630-28635

48. Hayashi, S., Rubinfeld, B., Souza, B., **Polakis, P.**, Wieschaus, E., and Levine, A. 1997 A Drosophila homolog of the tumor suppressor adenomatous polyposis coli down-regulates β -catenin but its zygotic expression is not essential for the regulation of armadillo. **Proc. Natl. Acad. Sci.** 94, 242-247.

49. Vleminckx, K., Rubinfeld, B., **Polakis, P.** and Gumbiner, B. 1997 The APC tumor suppressor protein induces a new axis in Xenopus embryos. **J. Cell. Biol.** 136, 411-420.

50. Rubinfeld, B., Robbins, P., El-Gamil, M., Albert, I., Porfiri, P. and **Polakis, P.** 1997 Stabilization of β-catenin by genetic defects in melanoma cell lines. **Science** 275, 1790-1792.

51. **Polakis, P.** The adenomatous polyposis coli (APC) tumor suppressor. 1997 **Biochem. Biophys. Acta**, 1332, F127-F147.

52. Rubinfeld, B., Albert, I., Porfiri, E., Munemitsu, S., and **Polakis, P** 1997 Loss of β-catenin regulation by the APC tumor suppressor protein correlates with loss of structure due to common somatic mutations of the gene. **Cancer Res.** 57, 4624-4630.

53. Porfiri, E., Rubinfeld, B., Albert, I., Hovanes. K., Waterman, M., and **Polakis, P.** 1997 Induction of a β-catenin-LEF-1 complex by wnt-1 and transforming mutants of β-catenin. **Oncogene** 15, 2833-2839.

54. Thomas JE, Smith M, Tonkinson JL, Rubinfeld B, and **Polakis P.**, 1997 Induction of phosphorylation on BRCA1 during the cell cycle and after DNA damage. **Cell Growth Differ.** 8, 801-809.

55. Hart, M., de los Santos, R., Albert, I., Rubinfeld, B., and **Polakis P.**, 1998 Down regulation of β-catenin by human Axin and its association with the adenomatous polyposis coli (APC) tumor suppressor, β-catenin and glycogen synthase kinase 3β. **Current Biology 8,** 573-581.

56. Polakis, P. 1998 The oncogenic activation of β-catenin. **Current Opinions in Genetics and Development** 9, 15-21

57. Matt Hart, Jean-Paul Concordet, Irina Lassot, Iris Albert, Rico del los Santos, Herve Durand, Christine Perret, Bonnee Rubinfled, Florence Margottin, Richard Benarous and **Paul Polakis**. 1999 The F-box protein β–TrCP associates with

phosphorylated β–catenin and regulates its activity in the cell. **Current Biology** 9, 207-10.

58. Howard C. Crawford, Barbara M. Fingleton, Bonnee Rubinfeld, **Paul Polakis** and Lynn M. Matrisian 1999 The metalloproteinase matrilysin is a target of β–catenin transactivation in intestinal tumours. **Oncogene** 18, 2883-91.

59. Meng J, Glick JL, **Polakis P**, Casey PJ. 1999 Functional interaction between Galpha(z) and Rap1GAP suggests a novel form of cellular cross-talk. **J Biol Chem**. 17, 36663-9

60. Vijayasurian Easwaran, Virginia Song, **Paul Polakis** and Steve Byers 1999 The ubiquitin-proteosome pathway and serine kinase activity modulate APC mediated regulation of β-catenin-LEF signaling. **J. Biol. Chem**. 274(23):16641-5.

61 Polakis P, Hart M and Rubinfeld B. 1999 Defects in the regulation of beta-catenin
in colorectal cancer. **Adv Exp Med Biol**. 470, 23-32

62 Shen Z, Batzer A, Koehler JA, **Polakis P**, Schlessinger J, Lydon NB, Moran MF. 1999 Evidence for SH3 domain directed binding and phosphorylation of Sam68 by Src. **Oncogene**. 18, 4647-53

64. Thomas GM, Frame S, Goedert M, Nathke I, **Polakis P**, Cohen P. 1999 A GSK3- binding peptide from FRAT1 selectively inhibits the GSK3-catalysed phosphorylation of axin and beta-catenin. **FEBS Lett**. 458, 247-51.

65. Peifer M, **Polakis P**. 2000 Wnt signaling in oncogenesis and embryogenesis--a look outside the nucleus. **Science** 287,1606-9.

66. **Polakis P.** 2000 Wnt signaling and cancer. **Genes Dev**;14, 1837-1851.

67. Spink KE, **Polakis P**, Weis WI 2000 Structural basis of the Axin-adenomatous polyposis coli interaction. **EMBO J** 19, 2270-2279.

68. Szeto , W., Jiang, W., Tice, D.A., Rubinfeld, B., Hollingshead, P.G., Fong, S.E., Dugger, D.L., Pham, T., Yansura, D.E., Wong, T.A., Grimaldi, J.C., Corpuz, R.T., Singh J.S., Frantz, G.D., Devaux, B., Crowley, C.W., Schwall, R.H., Eberhard, D.A., Rastelli, L., **Polakis, P.** and Pennica, D. 2001 Overexpression of the Retinoic Acid-Responsive Gene Stra6 in Human Cancers and its Synergistic Induction by Wnt-1 and Retinoic Acid. **Cancer Res** 61, 4197-4204.

69. Rubinfeld B, Tice DA, **Polakis P.** 2001 Axin dependent phosphorylation of the adenomatous polyposis coli protein mediated by casein kinase 1 epsilon. **J Biol Chem**

276, 39037-39045.

70. **Polakis P.** 2001 More than one way to skin a catenin. **Cell** 2001 105, 563-566.

71. Tice DA, Soloviev I, **Polakis P.** 2002 Activation of the Wnt Pathway Interferes withSerum Response Element-driven Transcription of Immediate Early Genes. **J Biol. Chem**. 277, 6118-6123.

72. Tice DA, Szeto W, Soloviev I, Rubinfeld B, Fong SE, Dugger DL, Winer J, Williams PM, Wieand D, Smith V, Schwall RH, Pennnica D, **Polakis P.** 2002 Synergistic activation of tumor antigens by wnt-1 signaling and retinoic acid revealed by gene expression profiling. **J Biol Chem**. 277,14329-14335.

73. **Polakis, P.** 2002 Casein kinase I: A wnt'er of disconnect. **Curr. Biol**. 12, R499.

74. Mao,W. , Luis, E., Ross, S., Silva, J., Tan, C., Crowley, C., Chui, C., Franz, G., Senter, P., Koeppen, H., **Polakis, P.** 2004 EphB2 as a therapeutic antibody drug target for the treatment of colorectal cancer. **Cancer Res**. 64, 781-788.

75. Shibamoto, S., Winer, J., Williams, M., Polakis, P. 2003 A Blockade in Wnt signaling is activated following the differentiation of F9 teratocarcinoma cells. **Exp. Cell Res.** 29211-20.

76. Zhang Y, Eberhard DA, Frantz GD, Dowd P, Wu TD, Zhou Y, Watanabe C, Luoh SM, **Polakis P,** Hillan KJ, Wood WI, Zhang Z. 2004 GEPIS--quantitative gene expression profiling in normal and cancer tissues. **Bioinformatics**, April 8

Genomics, gene expression and DNA arrays

David J. Lockhart & Elizabeth A. Winzeler

Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, California 92121, USA

Experimental genomics in combination with the growing body of sequence information promise to revolutionize the way cells and cellular processes are studied. Information on genomic sequence can be used experimentally with high-density DNA arrays that allow complex mixtures of RNA and DNA to be interrogated in a parallel and quantitative fashion. DNA arrays can be used for many different purposes, most prominently to measure levels of gene expression (messenger RNA abundance) for tens of thousands of genes simultaneously. Measurements of gene expression and other applications of arrays embody much of what is implied by the term 'genomics'; they are broad in scope, large in scale, and take advantage of all available sequence information for experimental design and data interpretation in pursuit of biological understanding.

Biological and biomedical research is in the midst of a significant transition that is being driven by two primary factors: the massive increase in the amount of DNA sequence information and the development of technologies to exploit its use. Consequently, we find ourselves at a time when new types of experiments are possible, and observations, analyses and discoveries are being made on an unprecedented scale. Over the past few years, more than 30 organisms have had their genomes completely sequenced, with another 100 or so in progress (see www.tigr.org or genomes@ncbi.nlm.nih.gov for a list). At least partial sequence has been obtained for tens of thousands of mouse, rat and human genes, and the sequence of two entire human chromosomes (chromosomes 21 and 22) has been determined[1,2]. Within the year, a large proportion of the human genome will be deciphered, in both public and private efforts, and the complete sequence of the mouse and other animal and plant genomes will undoubtedly follow close behind. Unfortunately, the billions of bases of DNA sequence do not tell us what all the genes do, how cells work, how cells form organisms, what goes wrong in disease, how we age or how to develop a drug. This is where functional genomics comes into play. The purpose of genomics is to understand biology, not simply to identify the component parts, and the experimental and computational methods take advantage of as much sequence information as possible. In this sense, functional genomics is less a specific project or programme than it is a mindset and general approach to problems. The goal is not simply to provide a catalogue of all the genes and information about their functions, but to understand how the components work together to comprise functioning cells and organisms.

To take full advantage of the large and rapidly increasing body of sequence information, new technologies are required. Among the most powerful and versatile tools for genomics are high-density arrays of oligonucleotides or complementary DNAs. Nucleic acid arrays work by hybridization of labelled RNA or DNA in solution to DNA molecules attached at specific locations on a surface. The hybridization of a sample to an array is, in effect, a highly parallel search by each molecule for a matching partner on an 'affinity matrix', with the eventual pairings of molecules on the surface determined by the rules of molecular recognition. Arrays of nucleic acids have been used for biological experiments for many years[3–8]. Traditionally, the arrays consisted of fragments of DNA, often with unknown sequence, spotted on a porous membrane (usually nylon). The arrayed DNA fragments often came from cDNA, genomic DNA or plasmid libraries, and the hybridized material was often labelled with a radioactive group. Recently, the use of glass as a substrate and fluorescence for detection, together with the development of new technologies for synthesizing or depositing nucleic acids on glass slides at very high densities, have allowed the miniaturization of nucleic acid arrays with concomitant increases in experimental efficiency and information content[9–14] (Fig. 1).

While making arrays with more than several hundred elements was until recently a significant technical achievement, arrays with more than 250,000 different oligonucleotide probes or 10,000 different cDNAs per square centimetre can now be produced in significant numbers[15,16]. Although it is possible to synthesize or deposit DNA fragments of unknown sequence, the most common implementation is to design arrays based on specific sequence information, a process sometimes referred to as 'downloading the genome onto a chip' (Fig. 1). There are several variations on this basic technical theme: the hybridization reaction may be driven (for example, by an electric field)[17,18]; other detection methods[19] besides fluorescence can be used; and the surface may be made of materials other than glass such as plastic, silicon, gold, a gel or membrane, or may even be comprised of beads at the ends of fibre-optic bundles[20–22]. Nonetheless, the key elements of parallel hybridization to localized, surface-bound nucleic acid probes and subsequent counting of bound molecules are ubiquitous, and high-density arrays of nucleic acids on glass (often called DNA microarrays, oligonucleotide arrays, GeneChip arrays, or simply 'chips') and their biological uses will be the focus of this review.

Global gene expression experiments

One of the most important applications for arrays so far is the monitoring of gene expression (mRNA abundance). The collection of genes that are expressed or transcribed from genomic DNA, sometimes referred to as the expression

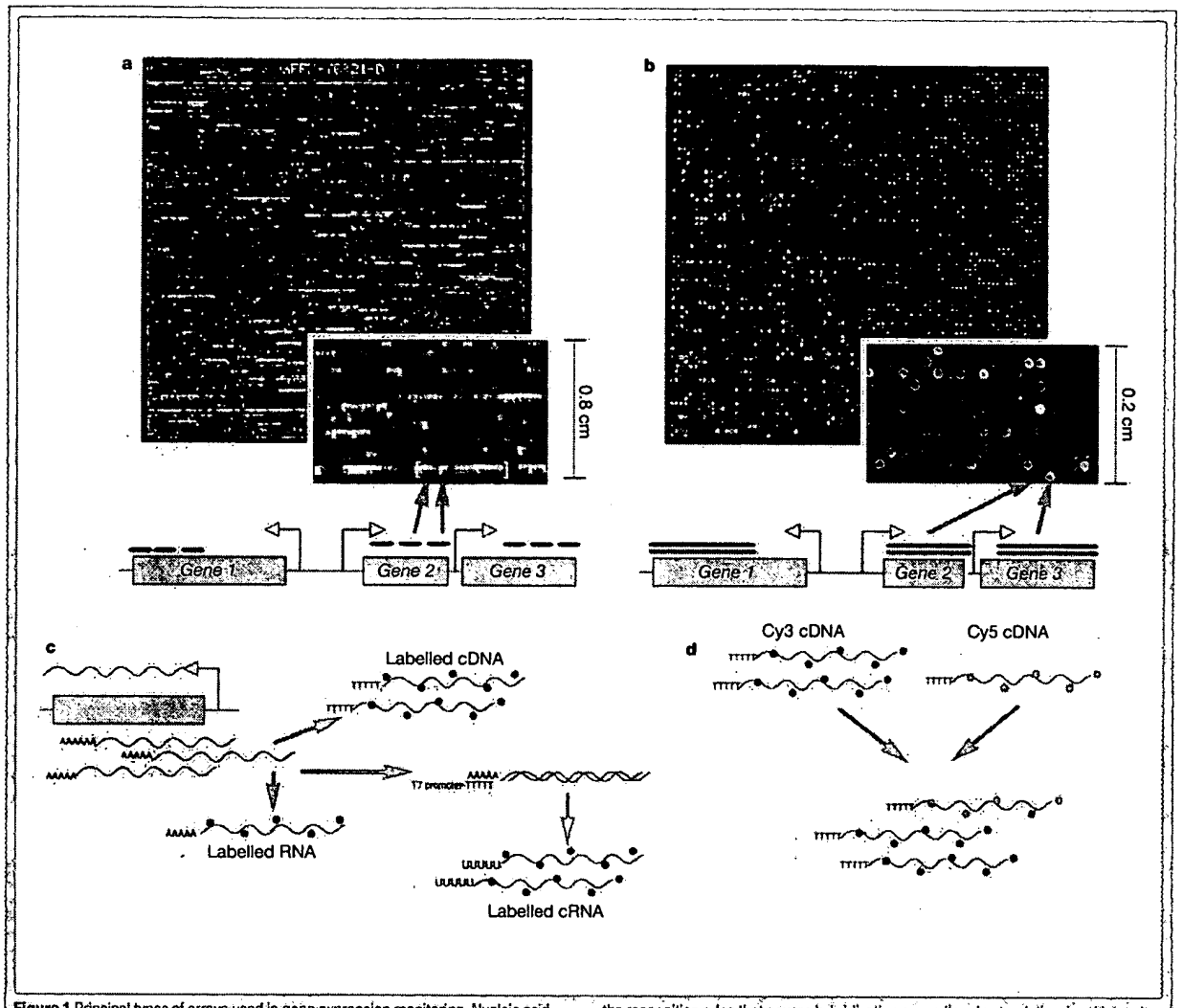

Figure 1 Principal types of arrays used in gene expression monitoring. Nucleic acid arrays are generally produced in one of two ways: by robotic deposition of nucleic acids (PCR products, plasmids or oligonucleotides) onto a glass slide[25] or *in situ* synthesis (using photolithography[15]) of oligonucleotides. Shown are pseudocolour images of **a**, an oligonucleotide array and **b**, a cDNA array after hybridization of labelled samples and fluorescence detection. In both cases the images have been coloured to indicate the relative number of yeast transcripts present under two different growth conditions (red, high in condition 1, low in condition 2; green, high in condition 2, low in condition 1; yellow, high under both conditions; black, low under both conditions). In the case of photolithographically synthesized arrays, ~10^7 copies of each selected oligonucleotide (usually 20 to 25 nucleotides in length) are synthesized base by base in hundreds of thousands of different 24 μm × 24 μm areas on a 1.28 cm × 1.28 cm glass surface. For robotic deposition, approximately one nanogram of material is deposited at intervals of 100–300 μm. Typically for oligonucleotide arrays, multiple probes per gene are placed on the array (20 pairs in the example shown here), while in the case of robotic deposition, a single, longer (up to 1,000 bp) double-stranded DNA probe is used for each gene or EST. In both cases, probes are usually designed from sequence located nearer to the 3′ end of the gene (near the poly-A tail in eukaryotic mRNA), and different probes can be used for different exons. After hybridization of labelled samples (typically overnight), the arrays are scanned and the quantitative fluorescence image along with the known identity of the probes is used to assess the 'presence' or 'absence' (more precisely, the detectability above thresholds based on background and noise levels) of a particular molecule (such as a transcript), and its relative abundance in one or more samples. Because the sequence of the oligonucleotide or cDNA at each physical location (or address) is generally known or can be determined, and because the recognition rules that govern hybridization are well understood, the signal intensity at each position gives not only a measure of the number of molecules bound, but also the likely identity of the molecules. Although oligonucleotide probes vary systematically in their hybridization efficiency, quantitative estimates of the number of transcripts per cell can be obtained directly by averaging the signal from multiple probes[15,26,30]. For technical reasons, the information obtained from spotted cDNA arrays gives the relative concentration (ratio) of a given transcript in two different samples (derived from competitive, two-colour hybridizations). Messenger RNAs present at a few copies (relative abundance of ~1:100,000 or less) to thousands of copies per mammalian cell can be detected[25,26,30], and changes as subtle as a factor of 1.3 to 2 can be reliably detected if replicate experiments are performed. **c**, Different methods for preparing labelled material for measurements of gene expression. The RNA can be labelled directly, using a psoralen–biotin derivative or by ligation to an RNA molecule carrying biotin[28]; labelled nucleotides can be incorporated into cDNA during or after reverse transcription of polyadenylated RNA; or cDNA can be generated that carries a T7 promoter at its 5′ end. In the last case, the double-stranded cDNA serves as template for a reverse transcription reaction in which labelled nucleotides are incorporated into cRNA. Commonly used labels include the fluorophores fluorescein, Cy3 (or Cy5), or nonfluorescent biotin, which is subsequently labelled by staining with a fluorescent streptavidin conjugate. **d**, Two-colour hybridization strategy often used with cDNA microarrays. cDNA from two different conditions is labelled with two different fluorescent dyes (usually Cy3 and Cy5), and the two samples are co-hybridized to an array. After washing, the array is scanned at two different wavelengths to detect the relative transcript abundance for each condition. cDNA array image courtesy of J. DeRisi and P. O. Brown (http://cmgm.stanford.edu/pbrown/yeastchip.html).

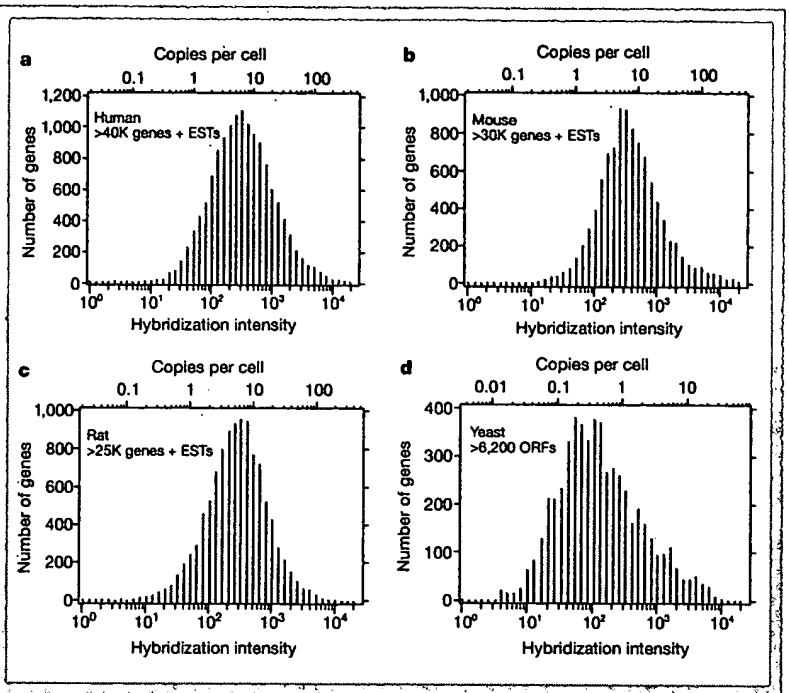

Figure 2 Messenger RNA abundance levels in different cells, tissues and organisms. **a**, Human HIV-infected T lymphocytes; **b**, mouse olfactory epithelium; **c**, rat brain; **d**, *S. cerevisiae* strain RY136 grown at 25 °C in rich medium. Levels of gene expression were measured using Affymetrix oligonucleotide arrays. For human, mouse and rat samples, hybridization intensities were converted to copies per cell (top axis) based on the signal from multiple control RNAs added to the samples at known concentrations. For yeast, the conversion was based on the signal from the TATA-binding protein (TBP) mRNA, which has been determined to be present at ~3.5 copies per cell when yeast cells are grown in rich medium[101]. Only those genes scored as 'present' are represented in the histograms. Data from multiple arrays containing probes for a different subset of genes and ESTs were combined to generate the plots for human (five arrays), mouse (five arrays) and rat (three arrays). All yeast ORFs were represented on a single array. For measurements that cover such a large number of genes, it is important to maintain high standards of data quality to keep false-positive results to a minimum. (For example, when monitoring 10,000 genes, even a low false-positive rate of 1% results in 100 false calls.) We find that the source of most false positives (in large part the result of setting the lowest possible thresholds in the interest of sensitivity) is random noise, biological variation, or the occasional array-specific physical defect, so observations made consistently in independent replicates yield a false-positive rate close to 0.01%, or only 1 in 10,000. In well controlled experiments involving specific biochemical, chemical and genetic perturbations, typically the number of expression differences is modest, with about 0.1–2% of the monitored genes changing by a factor of 1.8 or more, and only a small fraction of these changing by more than four- to fivefold[56–58,70–72,95,104]. For samples derived, for example, from different adult human or mouse tissues, or from normal versus advanced tumour tissue, the number of differences can be as large as 10–15% of the monitored genes[50–53]. The larger number of differences poses only minor difficulties for the technology, but analysis of the more complex results and the larger number of genes involved typically requires more sophisticated computational methods.

profile or the 'transcriptome', is a major determinant of cellular phenotype and function. The transcription of genomic DNA to produce mRNA is the first step in the process of protein synthesis, and differences in gene expression are responsible for both morphological and phenotypic differences as well as indicative of cellular responses to environmental stimuli and perturbations. Unlike the genome, the transcriptome is highly dynamic and changes rapidly and dramatically in response to perturbations or even during normal cellular events such as DNA replication and cell division[23,24]. In terms of understanding the function of genes, knowing when, where and to what extent a gene is expressed is central to understanding the activity and biological roles of its encoded protein. In addition, changes in the multi-gene patterns of expression can provide clues about regulatory mechanisms and broader cellular functions and biochemical pathways. In the context of human health and treatment, the knowledge gained from these types of measurements can help determine the causes and consequences of disease, how drugs and drug candidates work in cells and organisms, and what gene products might have therapeutic uses themselves or may be appropriate targets for therapeutic intervention.

Past discussions of arrays have often centred on technical issues and specific performance characteristics[25]. Now that nucleic acid arrays have been constructed for many different organisms[14,26–29] and used successfully to measure transcript abundance in a host of different experiments, the focus of interest has thankfully shifted. Investigators are now more concerned with questions concerning experimental design, data analysis, the use of small amounts of mRNA from limited sources, the best ways to extract biological meaning from the results, pathway and cell-circuitry modelling, and medical uses of expression patterns.

Array-based gene expression monitoring

One way to think of measurements with arrays is that they are simply a more powerful substitute for conventional methods of evaluating mRNA abundance. For some early experiments, only a relatively small set of genes, which were thought to be important to a process, were included on the arrays[12,30]. However, such experiments did not capitalize on the arrays' potential: a key advantage of using arrays, especially those that contain probes for tens of thousands of different genes, is that it is not necessary to guess what the important genes or mechanisms are in advance. Instead of looking only under the proverbial lamppost, a broader, more complete and less biased view of the cellular response is obtained (Figs 2, 3).

The breadth of array-based observations almost guarantees that surprising findings will be made. A recent study measured the transcriptional changes that occur as cells progress through the normal cell-division cycle in humans for approximately 40,000 genes (R. J. Cho *et al.*, unpublished results). In addition to the induction of DNA replication genes and genes involved with cell-cycle control and chromosome segregation that would be expected at specific stages in the cell cycle, a large collection of genes involved with smooth muscle function, apoptosis and intercellular adhesion and cell motility were found to be upregulated during a specific phase. The expected results act effectively as internal controls that provide a certain amount of validation (and comfort), while new information is obtained by a systematic search of a larger part of 'gene space'. In addition, because arrays often contain probes for genes of unknown function (and often with only partial sequence information), any outcome for these could be considered, in some sense, both surprising and novel (although clearly requiring further characterization).

Other gene expression methods

Not surprisingly, there are other ways to measure mRNA abundance, gene expression and changes in gene expression. For measuring gene expression at the level of mRNA, northern blots, polymerase chain reaction after reverse transcription of RNA (RT-PCR), nuclease

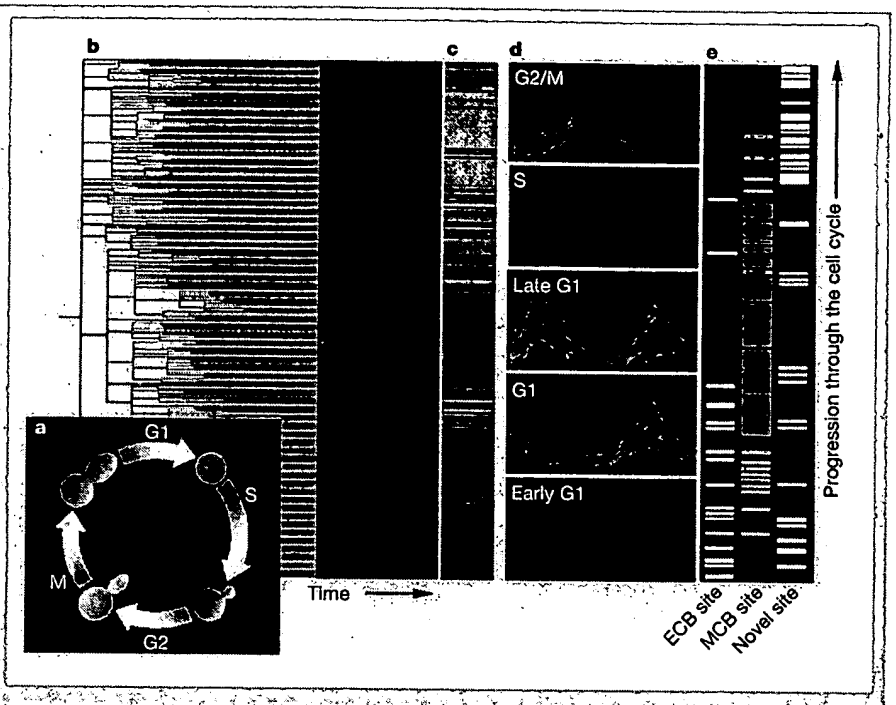

Figure 3 Methods for analysing gene expression data shown for measurements of expression in the cell cycle of *S. cerevisiae*. **a**, Yeast cells were synchronized and cells were collected every ten minutes throughout two complete synchronous cycles (18 time points in total are shown). Expression data were collected by hybridizing labelled cDNA samples to high-density oligonucleotide arrays. Transcript levels were determined for almost every gene in the genome for every time point[24]. A sample of 409 genes (from a total of 6,000) that showed both a significant (more than twofold) fluctuation in transcript levels during the time course and cell cycle-dependent periodicity were selected for further analysis. **b**, Dendrogram indicating similarity of expression profiles, calculated using the Pearson correlation function in the GeneSpring software package (Silicon Genetics, San Carlos, CA). For display purposes, the relative expression levels were plotted in red (high) and blue (low). **c**, The genes were divided into five different temporal expression classes (red, early G1; light blue, G1; green, late G1; dark blue, S; orange, G2/M) using K-tuple means clustering (also using GeneSpring software) and the clusters were named according to their time of peak expression within the cell cycle. **d**, Line graphs for all genes in the clusters defined in **b**. **e**, Location of cell cycle-regulated genes within the dendrogram in a that have *cis*-regulatory sequence elements in the 500 bp upstream of their promoter. Column 1, MCB sites (ACGCGT); column 2, ECB sites (TTWCCCNNNNAGGAA); column 3, a new sequence (GTAAACAA or TTGTTTAC) was identified that was statistically associated ($p = 1.77 \times 10^{-7}$ for the forward direction, $p = 0.003$ for the reverse) with the promoter regions of genes whose expression peaked in G2/M phase.

protection, cDNA sequencing, clone hybridization, differential display[31], subtractive hybridization, cDNA fragment fingerprinting[32–35] and serial analysis of gene expression (SAGE)[36] have all been put to good use to measure the expression levels of specific genes, characterize global expression profiles or to screen for significant differences in mRNA abundance. But if messenger RNA is only an intermediate on the way to production of the functional protein products, why measure mRNA at all? One reason is simply that protein-based approaches are generally more difficult, less sensitive and have a lower throughput than RNA-based ones. But more importantly, mRNA levels are immensely informative about cell state and the activity of genes, and for most genes, changes in mRNA abundance are related to changes in protein abundance. Because of its importance, however, many methods have been developed for monitoring protein levels either directly or indirectly (see review in this issue by Pandey and Mann, pages 837–846). These include western blots, two-dimensional gels, methods based on protein or peptide chromatographic separation and mass spectrometric detection[37–40], methods that use specific protein-fusion reporter constructs and colorimetric readouts[41–44], and methods based on characterization of actively translated, polysomal mRNA[45–47].

The importance of the protein-based methods is that they measure the final expression product rather than an intermediate. In addition, some of them enable the detection of post-translational protein modifications (for example, phosphorylation and glycosylation) and protein complexes, and in some cases, yield information about protein localization, none of which are obtained directly by measurements of mRNA. There is no question that protein- and RNA-based measurements are complementary, and that protein-based methods are important as they measure observables that are not readily detected in other ways.

Human disease, gene expression and discovery

Genomics and gene expression experiments are sometimes derided as 'fishing expeditions'. Our view is that there is nothing wrong with a fishing expedition[48] if what you are after is 'fish', such as new genes involved in a pathway, potential drug targets or expression markers that can be used in a predictive or diagnostic fashion. Because the arrays can be designed and made on the basis of only partial sequence information, it is possible to include genes in a survey that are completely uncharacterized. In many ways, the spirit of this approach is more akin to that of classical genetics in which mutations are made broadly and at random (not only in specific genes), and screens or selections are set up to discover mutants with an interesting phenotype, which then leads to further characterization of specific genes.

Such broad discovery experiments are probably better described as 'question-driven' rather than hypothesis-driven in the conventional sense. But that is not to diminish their value for understanding basic biological processes and even for understanding and treating human disease. For example, by analysing multiple samples obtained from individuals with and without acute leukaemia or diffuse large B-cell lymphoma, gene expression (mRNA) markers were discovered that could be used in the classification of these cancers[49,50]. The importance of monitoring a large number of genes was well illustrated in these studies. Golub *et al.*[49] found that reliable predictions could not be made based on any single gene, but that predictions based on the expression levels of 50 genes (selected from the more than 6,000 monitored on the arrays) were highly accurate. The results of both of these studies indicate that measurements with more individuals and more genes will be needed to identify robust expression markers that are predictive of clinical outcome. But even with the limited initial data it was possible to help clarify an unusual case (classic leukaemia presentation but atypical morphology) and to use this information to guide the patient's clinical care.

It is also possible to take a related approach to help understand what goes wrong in cancerous, transformed cells and to identify the genes responsible for disease. Causative effects and potential

therapeutic targets can be identified by determining which genes are upregulated in different tumour types[51–55], and specific candidate genes can be intentionally overexpressed in cell lines or cells treated with growth factors in order to identify downstream target genes and to explore signalling pathways[56–58]. Tumorigenesis is often accompanied by changes in chromosomal DNA, such as genetic rearrangements, amplifications or losses of particular chromosomal loci, and developmental abnormalities, such as Down's or Turner's syndrome, may arise from aberrations in DNA copy number. Because genomic DNA can be interrogated in much the same way as mRNA, comparisons of the copy number of genomic regions or the genotype of genetic markers can be used to detect chromosomal regions and genes that are amplified or deleted in cancerous or pre-cancerous cells. By using arrays containing probes for a large number of genes or polymorphic markers, changes in DNA copy number have been detected in both breast cancer cell lines and in tumours[59–61]. The identification of when and where changes in copy number or chromosomal rearrangements have occurred can be used in both the classification of cancer types and the identification of regions that may harbour tumour-suppressor genes.

Whole-genome hypotheses

The use of genomics tools such as arrays does not, of course, preclude hypothesis-driven research. For fully sequenced organisms, arrays containing probes for every annotated gene in the genome have been produced[14,26]. With these one can ask, for example, whether a transcription factor has a global role in transcription (affecting all genes) or a specific role (affecting only some). Holstege *et al.*[62] used this type of application in a genome-wide expression analysis in yeast to functionally dissect the machinery of transcription initiation. Similarly, genes located near the ends of chromosomes in yeast (as well as genes at the mating-type locus) are known to be transcriptionally 'silent'. Full genome arrays allow the chromosomal landscape of silencing to be mapped, and make it possible to test whether what is true for a handful of well-studied genes near the telomeres is true for all telomeric genes, and whether any centromere-proximal genes are also transcriptionally silenced[63].

It is important to emphasize that these new, parallel approaches do not replace conventional methods. Standard methods such as northern blots, western blots or RT-PCR are simply used in a more targeted fashion to complement the broader measurements and to follow-up on the genes, pathways and mechanisms implicated by the array results. Because the incidence of false-positive results can be made sufficiently low (see Fig. 2), it is not necessary to independently confirm every change for the results to be valid and trustworthy, especially if conclusions are based on changes in sets of genes rather than individual genes. More detailed follow-up is recommended if a gene is being chosen, for example, as a drug target, as a candidate for population genetics studies, or as the target for the construction of a knockout mouse.

Does gene expression indicate function?

As additional, uncharacterized open reading frames (ORFs) are identified in different organisms by the various genome sequencing projects, researchers have begun to ask whether the expression pattern for a gene can be used to predict the functional role of its protein product. An increasingly common approach involves using the gene expression behaviour observed over multiple experiments to first cluster genes together into groups (see Fig. 3), either by manual examination of the data[24], or by using statistical methods such as self-organizing maps[64], K-tuple means clustering or hierarchical clustering[23,65,66]. The basic assumption underlying this approach is that genes with similar expression behaviour (for example, increasing and decreasing together under similar circumstances) are likely to be related functionally. In this way, genes without previous functional assignments can be given tentative assignments or assigned a role in a biological process based on the known functions of genes in the same expression cluster (that is, the concept of 'guilt-by-association'). The validity of this approach has been demonstrated for many genes in *Saccharomyces cerevisiae*, a simple organism for which the entire genomic sequence and the functional roles of approximately 60% of the genes are known[24,65,67] (Fig. 4). Although not logically rigorous, the utility of the guilt-by-association approach has been demonstrated, as genes already known to be related do, in fact, tend to cluster together based on their experimentally determined expression patterns (Fig. 4). The approach is made more systematic and statistically sound by calculating the probability that the observed functional distribution of differentially expressed genes could have happened by chance. The application of statistical rigour is essential to avoid overly subjective interpretations of the results based on the predispositions, prior knowledge and interests of the individual researcher.

A tentative functional assignment may not be much more than a low-resolution description or general classification. Descriptions of this type are similar to those that come out of more classical genetic screens and selections, which have provided the vast majority of functional annotations to date — they indicate that genes are involved with a particular cellular phenotype and that they are likely

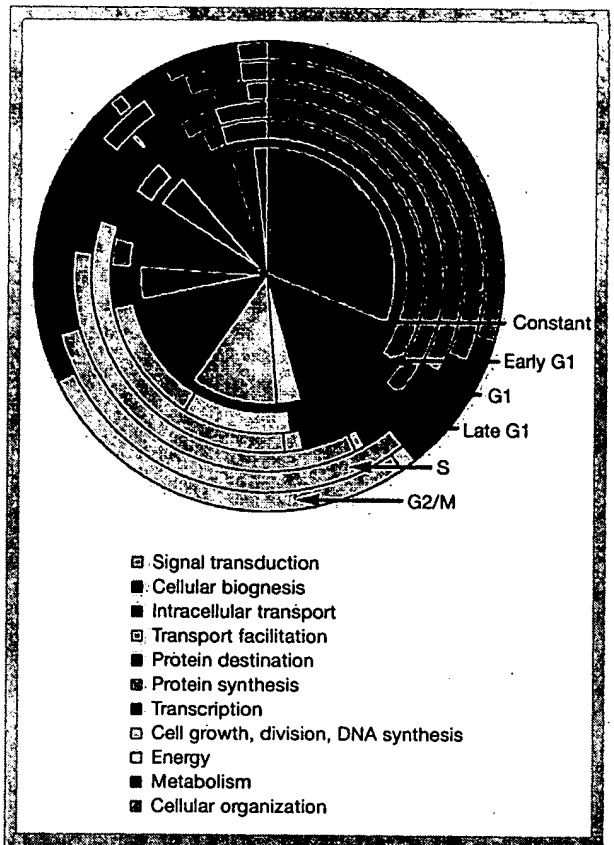

Figure 4 The 'guilt-by-association' method for assigning gene function. Functional distribution (using categories from MIPS: http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/index.html) of yeast genes whose periodic expression peaked at different times in the yeast cell cycle (outer rings) or was constant throughout the cell cycle (inner circle)[24]. A much larger fraction of cell cycle-modulated genes is important in DNA synthesis, cell growth or cell division. Although there is a strong correlation between distinct expression profiles and functional assignments, specific expression behaviour should not be taken as sufficient evidence for functional assignment: not all genes involved in DNA replication are expressed periodically in the cell cycle, and some genes that do not need to be cell cycle-regulated are transcribed in a periodic fashion.

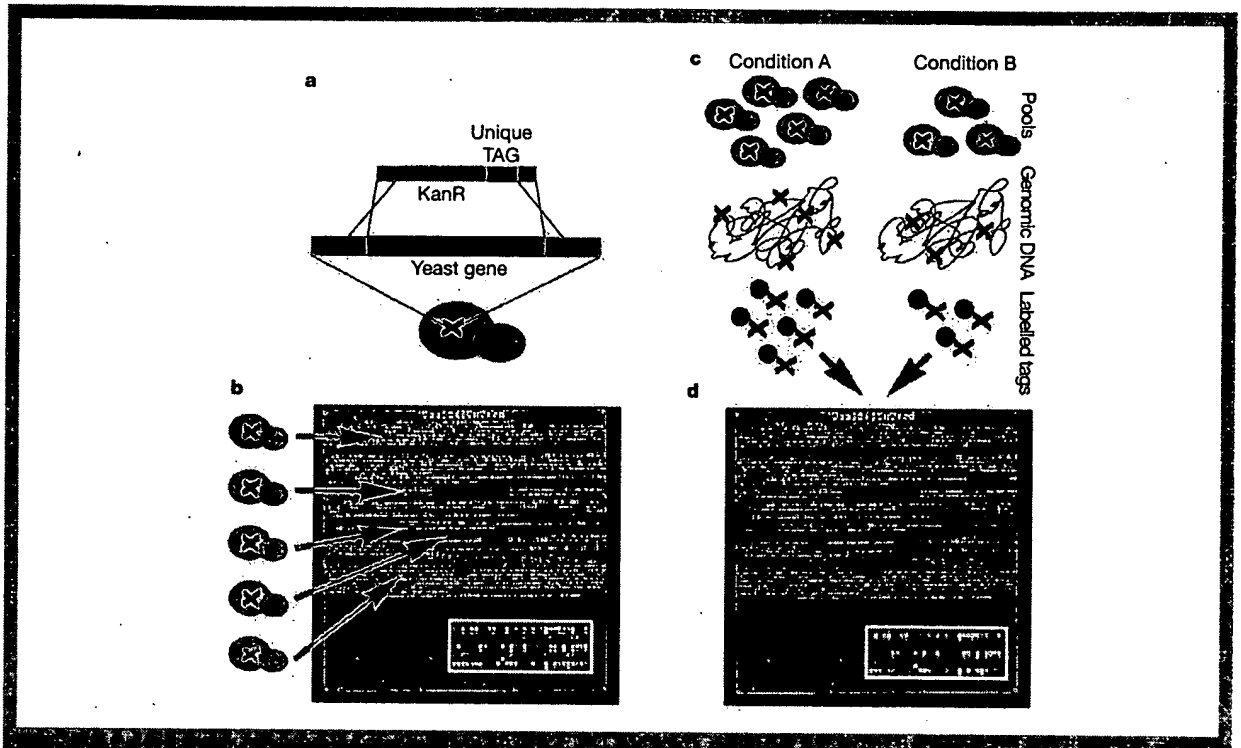

Figure 5 Generic oligonucleotide tag arrays for parallel phenotyping of mutant yeast strains. a, Many *S. cerevisiae* strains, each carrying a specific deletion of one of the more than 6,000 ORFs in the yeast genome, have been constructed[64] by replacing individual genes with an antibiotic resistance cassette and a unique gene-specific 20-mer 'barcode', represented by an X. b, The barcode for each deletion strain corresponds to a specific location on an array that contains oligonucleotide probes that are complementary to the barcode sequences. c, Pools of different yeast strains can be assembled and grown under different conditions. After competitive growth, PCR is used to amplify the barcodes from genomic DNA isolated from the pools; the PCR products are subsequently labelled. d, By comparing the hybridization patterns of two different pools (before and after treatment with a drug, for example), the fitness of the strains can be assessed quantitatively. In this case, yeast genes required for sporulation or germination are represented in red, whereas yeast genes that are unnecessary for the process are shown in yellow. These same 20-mer sequences and the accompanying arrays are generic in design, and can be used to read the results of different types of 'bar-coded' reactions, such as those used for genotyping of human polymorphic loci[165]. Images provided by R. M. Williams and R. W. Davis.

to be involved with a certain set of other genes and processes. This allows researchers to focus attention on a smaller subset of genes, many of which may not have been obvious candidates in the absence of the global expression observations. This overall approach highlights the importance of functional annotation and careful curation of existing sequence, function and knowledge databases (see below). Expression results covering thousands or even tens of thousands of genes and expressed sequence tags (ESTs) will be only partly interpretable given the functional and biological information available at the time they are initially generated. Our ability to extract knowledge from measurements of global gene expression tends to increase with time as additional information becomes available, and results can be subjected to further interrogation in the light of new information, observations, questions and hypotheses.

Gene expression and the regulation of transcription

When information on the complete genome sequence is available, as is the case for increasing numbers of small and even larger genomes, gene expression data can be used to identify new *cis*-regulatory elements (genomic sequence motifs that are over-represented in the genomic DNA in the vicinity of similarly behaving genes) and 'regulons' (sets of co-regulated genes), the basic units of the underlying cellular circuitry (Fig. 3d). In fact, the correlation between the presence of specific sequence motifs in promoter regions and gene expression patterns may be stronger than the correlation between functional categories and gene expression patterns. In yeast studies, more than 50% of the genes that are transcribed in a cell cycle-specific manner and whose transcript abundance peaks in the G1 phase of the cell cycle have an MCB (Mlu cell-cycle box) within 500 base pairs (bp) of their translational start site[24,68,69]. Similar observations have been made for yeast genes whose transcription is induced during sporulation[67]. In addition, new *cis*-regulatory elements may be revealed by examining classes of co-regulated genes (Fig. 3d). With sufficiently large numbers of experimental observations of expression behaviour, the boundaries and all functioning sequence variants of *cis*-regulatory elements might be predicted without the need for the more conventional approach using site-directed mutagenesis ('promoter bashing'). The expression-based method will be especially valuable in exotic organisms, such as *Plasmodium falciparum*, the causative agent for malaria, for which experimental identification or verification of transcription factor binding sites is difficult.

Gene expression profiles as 'fingerprints'

An often overlooked aspect of measurements of global gene expression is that the sequence or even the origin of the arrayed probes does not need to be known to make interesting observations — the complex profiles, consisting of thousands of individual observations, can serve as transcriptional 'fingerprints'. The fingerprints can be used for classification purposes or as tests for relatedness, in a similar manner to the way in which DNA fingerprints are used in paternity testing. In one example, transcriptional fingerprints have been used to determine the target of a drug[70]. The basic idea is that if a drug interacts with and inactivates a specific cellular protein, the phenotype of the drug-treated cell should be very similar to the phenotype

of a cell in which the gene encoding the protein has been genetically inactivated, usually through mutation. Thus, by comparing the expression profile of a drug-treated cell to the profiles of cells in which single genes have been individually inactivated, specific mutants can be matched to specific drugs, and therefore, targets to drugs. In a demonstration of this concept, the gene product of the *his3* gene was identified correctly as the target of 3-aminotriazole[70]. Similarly, profiles have been used in the classification of cancers and the classification schemes did not depend on any specific information about the genes involved[49,50], although that information can be used to draw further biological and mechanistic conclusions. Finally, expression profiles can be used to classify drugs and their mode of action. For example, the functional similarity and specificity of different purine analogues have been determined by comparing the genome-wide effects on treated yeast, murine and human cells[71,72].

Expression measurements from small amounts of RNA

An important frontier in the development of gene expression technology involves reduction of the required amount of starting material. Most array-based expression measurements are done using RNA from a million or more cells, and obtaining such a relatively large sample is not a problem in many types of studies (for example, litres of yeast cells can be grown easily). However, in some cases, it is important or even necessary to use fewer cells, as when using a small organ from a fly or worm, sorted cells that express a rare marker, or laser-capture microdissected[73–75] tumour tissue. Efficient and reproducible mRNA amplification methods are required, and there are two primary approaches that show significant promise. The first is a PCR-based approach that has been used to make single-cell cDNA libraries[76–78]. We have found that the amplification is efficient and reproducible, but that the relative abundance of the cDNA products is not well correlated with the original mRNA levels (D. Giang and D. J. Lockhart, unpublished results), although normalization and referencing strategies can be used (D. de Graaf and E. Lander, personal communication).

The second approach avoids PCR altogether and uses multiple rounds of linear amplification based on cDNA synthesis and a template-directed *in vitro* transcription (IVT) reaction[79–81]. This method has been used to characterize mRNA from single live neurons[81] and even subcellular regions, and more recently to amplify mRNA from 500 to 1,000 cells from microdissected brain tissues for hybridization to spotted cDNA arrays[82]. We have found that the multiple-round cDNA/IVT amplification method produces sufficient quantities of labelled material starting with as little as 1–50 ng total RNA, is highly reproducible (correlation coefficients greater than 0.97), and introduces much less quantitative bias than PCR-based amplification (D. Giang and D. J. Lockhart, unpublished results). These amplification methods facilitate the possibility of monitoring large number of genes starting with very limited amounts of RNA and very few cells. The combination of arrays and powerful amplification strategies promises to be especially important for studies that use human biopsy material from inhomogeneous tissue, and in the areas of developmental biology, immunology and neurobiology.

Genome analysis using arrays

Although nucleic acid arrays are often equated with gene expression analysis, they may be used to collect much of the data that are obtained presently by Southern or northern blot hybridization techniques, but in a more highly parallel fashion (Figs 5, 6). Their utility in polymorphism detection and genotyping is described elsewhere (see review in this issue by Roses, pages 857–865), but there are many additional uses for these versatile tools. For example, genomic DNA samples can be manipulated experimentally to select for particular regions before hybridization to obtain specific types of information. In yeast, the location of hundreds of chromosomal origins of replication can be determined in parallel by enriching for early-replicating regions using a variation of the Meselsohn–Stahl procedure and then hybridizing the resulting DNA to full genome arrays (E. A. Winzeler *et al.*, unpublished results). Similarly, as probes for more intergenic regions are synthesized on arrays, it becomes possible to identify protein-binding sites: fragmented chromatin can be crosslinked to a protein and then immunoprecipitated with an antibody to that protein. The DNA fraction of the immunoprecipitate can be labelled and hybridized to identify the approximate location of the binding site. In addition, full genome arrays can be used in the analysis of plasmid libraries in genetic selections such as two-hybrid screens[83] or, in principal, for any other type of experiment in which the information is contained in the form of RNA or DNA. Arrays also have applications in biophysical chemistry and biochemistry. For example, single-stranded DNA arrays were converted enzymatically into arrays of double-stranded DNA to characterize the interactions of proteins, and potentially other types of molecules, with double-stranded DNA[84].

Gene expression and cell circuitry

Is it reasonable to consider the cell as a complex analogue circuit, and to attempt to reverse-engineer the cell circuitry much like an electrical engineer would do by measuring currents and voltages at a variety of nodes and under a variety of input conditions? In the case of the cell, expression levels and expression changes might take the place of electrical measurements, and could be measured under many experimental conditions. Is it possible that a genetic or cellular circuit of reasonable complexity could be adequately decoded or modelled, and if so, how many and what types of measurements and perturbations (or 'inputs') would be required so that the problem was not hopelessly underdetermined[85–89]? Reasonably detailed circuit diagrams can be drawn and simulations of simple genetic circuits have been performed for systems of low complexity (for example, the lytic cycle of phage lambda, and simple control networks in *Escherichia coli* bacteria[90]). But the situation is considerably more complex in the case of a eukaryotic cell. Using yeast as an example, if we assume that the expression level for each gene can be one of only four levels (off, low, medium or high), then if the 6,200 yeast genes behave independently, there are $6{,}200^4$, or $\sim 1.5 \times 10^{15}$ possible expression states. Of course, the expression levels of different genes are not all independent of one another, and there are some states that are physically unrealistic (for example, all genes 'off' or all genes 'high'), but the number of possible cellular configurations is very large. In addition, coupling between circuit components, the effects of nonlinear feedback, redundancy and even noise and stochastic events make simulating a circuit of this complexity a rather daunting task, and not all relationships and cellular events are reflected at the level of mRNA abundance.

Least clear may be what types of perturbations or inputs are likely to be the most informative in terms of defining the relationships between genes and pathways, and what might be a minimal set of 'orthogonal perturbations' (treatments, genetic manipulations or growth conditions that have minimal overlap in their direct cellular effects). Certainly it is possible to delete every yeast gene one at a time (or even several at a time) and measure the expression profile for each mutant strain under a set of different growth conditions[70,91]. It is also possible to grow yeast on a matrix of thousands of different conditions and measure the resulting expression profiles for a range of mutated strains. It is clear that extensive experiments of this type, combined with information from other measurements such as yeast two-hybrid protein–protein interaction screens[92], and measurements of protein levels, modification states and cellular localization will lead to useful groupings of genes in terms of function and regulation (that is, a genetic, molecular and functional taxonomy), and to supply some reasonably detailed information about the relationships between certain genes and pathways. In addition, sets of perturbations directed towards specific functions and cellular processes will allow higher-resolution and even mechanistic

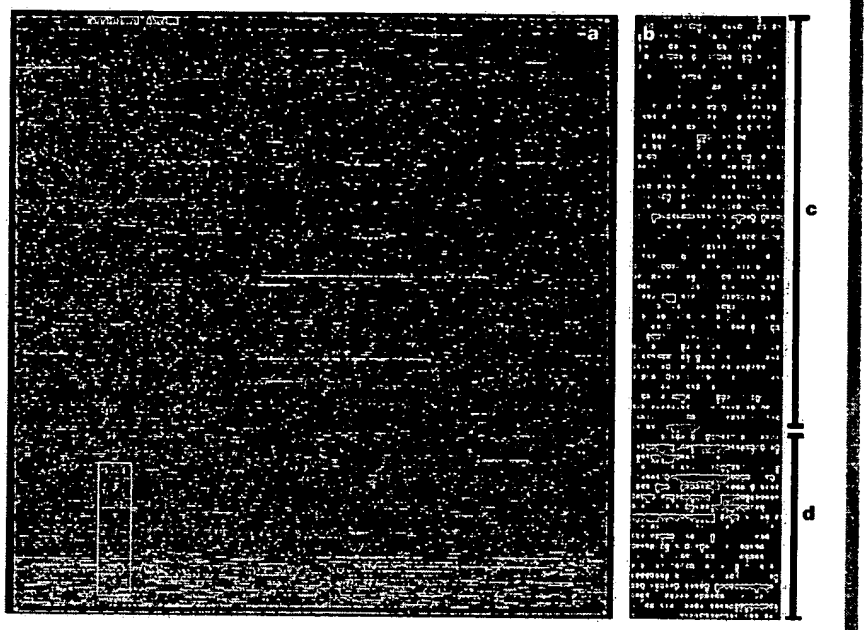

Figure 6 Comparative genome hybridization using arrays[100,101]. a, Two arrays containing probes to yeast (the complete genome sequence of *S. cerevisiae* strain S288c and some *S. cerevisiae* DNA not present in S288c) were hybridized with fragmented, labelled genomic DNA from two different yeast strains commonly used in genetic studies (W303 and SK1). Red indicates the location of probes that hybridize efficiently only to DNA from the W303 strain, green indicates probes that hybridize only to SK1 DNA, and yellow indicates probes that hybridize equally to the DNA from both strains. b, Enlargement of the boxed region in a. c, Region of the array containing probes to relatively unique protein-coding regions of the genome. d, Probes to non-unique regions of the genome (transposable elements, telomeric sequences, transfer RNAs and ribosomal RNAs). Genome regions that are present, absent, or found at higher or lower copy numbers in the two strains are readily detected. The large amount of allelic variation between the strains can be used in mapping studies[100]. Related approaches can be used in typing microbial isolates[99,102] or to identify genetic abnormalities in tumours.

information for significant parts of the overall circuitry[62,93]. However, given the tremendous complexity of the system, it is unlikely that a complete and detailed cellular circuit diagram will result for even single-celled eukaryotes such as yeast any time in the near future. But that is not to say that construction of even first-order global models and semi-quantitative circuit diagrams is not extremely useful. Such models serve to organize current information, relationships and hypotheses, and can be tremendously helpful for testing new hypotheses, interpreting new observations, designing new experiments and predicting the likely effects of particular chemical, genetic or cellular perturbations. They also serve as a scaffold upon which to build higher-resolution, more quantitative and complete models.

Can we have too much data?

Contrary to what is sometimes thought, the biggest problem for making sense of the extensive results from genomics experiments is not that there is too much data or that there are insufficiently sophisticated algorithms and software tools for querying and visualizing data on this scale. Larger problems of data management and analysis have been solved by airlines, financial institutions, global retailers, high-energy and plasma physicists, the military and global weather predictors, among others. It is often beneficial to have a large number of measurements[94] and sometimes more data make it possible to analyse results that might otherwise have been too 'messy', and to detect patterns and relationships that would not have been obvious or have sufficient statistical significance with smaller data sets. In many types of studies, it is not possible to control completely all variables, and the individual differences between common sample types may be significant because of experimental difficulties (for example, tissue inhomogeneity or variations in sample procedures) or individual genetic variation (for example, different patients or different tumours). But such factors do not preclude the discovery of some genes that clearly 'cluster' or differentiate between the sample sets. For example, meaningful results can be extracted from the analysis of human tissue collected at different hospitals, by different surgeons and at different times. An essential requirement in these types of studies is that a sufficient number of experiments be performed across multiple individuals and multiple tissue or tumour samples to account for individual variation and possible tissue inhomogeneity. Furthermore, confidence in the results is increased as conclusions are based on sets of genes that show a consistent response and that are consistently different between two or more sets of results[49,50,52,53,95].

Making sense of genomic results

Although the difficulties of sample collection, data collection and experimental design should not be underestimated, one of the most challenging aspects of gene expression analysis is making sense of the vast quantities of data and extracting conclusions and hypotheses that are biologically meaningful. From experiments on global gene expression, we may obtain data for thousands of genes, often forcing us to consider processes, functions and mechanisms about which we know very little. Thus, there is a need for more sophisticated systems of knowledge representation (or 'knowledge bases') that organize the data, facts, observations, relationships and even hypotheses that form the basis of our current scientific understanding. This information needs to be more than just stored; it needs to be available in a way that helps scientists understand and interpret the often complex observations that are becoming increasingly easy to make. Unfortunately, the fact is that the scientific literature has been somewhat haphazardly built, without the benefit of a controlled or restricted vocabulary and a well defined semantic and grammar. To take full advantage of the abilities of the new technologies and the rapidly increasing amount of sequence information it is absolutely essential to incorporate the facts, ideas, connections, observations and so forth, which exist in the scientific literature and in the minds of scientists, into a form that is systematic, organized, linked, visualized and searchable. This clearly requires a great deal of dedicated, systematic human effort, but progress has been made. Databases such as the *Saccharomyces* Genome Database (SGD: genome-www.stanford.edu/Saccharomyces), the Munich Information Center for Protein Sequences (MIPS:

www.mips.biochem.mpg.de), WormBase (www.wormbase.org), the Kyoto Encyclopedia of Genes and Genomes (KEGG: www.genome.ad.jp/kegg), the Encyclopedia of *E. coli* Genes and Metabolism (EcoCyc: http://ecocyc.panbio.com/ecocyc) and FlyBase (flybase.bio.Indiana.edu/) incorporate sequence, genetics, gene expression, homology, regulation, function and phenotype information in an organized and useable form[96–102]. But a step beyond databases of this type are ones in which concepts as well as facts are more fully integrated and related, allowing connections to be made between initially disparate observations and information, and across organisms. It is conceivable that the next step will evolve to the level of a biological 'expert system', not unlike the expert system ('Big Blue') that IBM scientists and engineers built to play chess (successfully) against the world's best chess player. Despite the potential for advancement on this front, it seems unlikely that computational tools will ever replace the trained human brain when it comes to making biological sense of new results. However, the appropriate tools are needed to bring information and relationships to scientist's fingertips so that the most insightful questions can be asked and the most meaningful interpretations made.

Conclusion

For these array-based methods to become truly revolutionary, they must become an integral part of the daily activities of the typical molecular biology laboratory. Despite their impressive and rapidly growing résumé, these technologies are still in their infancy, with plenty of room for technical improvements, further development, and more widespread acceptance and accessibility. We expect that the pattern of development and use of arrays and other parallel genomic methodologies will be similar to that seen for computers and other high-tech electronic devices, which started out as exotic and expensive tools in the hands of the few developers and early adopters, and then moved quickly to become easier to use, more available, less expensive and more powerful, both individually and because of their ubiquity. In fact, nucleic acid array-based methods that previously seemed exotic, and too expensive, are becoming routine as indicated by the huge increase in the number of publications that incorporate data obtained in this way. Despite the relative youth of these approaches, the achievement of technical goals that would have seemed like science fiction only a few years ago is now clearly in view. For example, we expect that measuring the expression level of essentially every gene (including variant splice forms) on an array or two starting with RNA from a small number of cells, or even a single cell, will soon be possible owing to advances in single-cell handling and RNA amplification methods, the output of large-scale sequencing efforts and achievable advances in array technology. In the future, arrays of peptides, proteins, small molecules, mRNAs, clones, tissues, cells and even multicellular organisms such as the nematode worm *Caenorhabditis elegans* may also become common. The combined use of all of these highly parallel methods, along with sequence information, computational tools, integrated knowledge databases, and the traditional approaches of biology, biochemistry, chemistry, physics, mathematics and genetics, increases the hopes of understanding the function and regulation of all genes and proteins, deciphering the underlying workings of the cell, determining the mechanisms of disease, and discovering ways to intervene with or prevent aberrant cellular processes in order to improve human health and well-being. □

1. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
2. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
3. Lennon, G. G. & Lehrach, H. Hybridization analyses of arrayed cDNA libraries. *Trends Genet.* **7**, 314–317 (1991).
4. Kafatos, F. C., Jones, C. W. & Efstratiadis, A. Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic Acids Res.* **7**, 1541–1552 (1979).
5. Gillespie, D. & Spiegelman, S. A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane. *J. Mol. Biol.* **12**, 829–842 (1965).
6. Southern, E. M. *et al.* Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. *Nucleic Acids Res.* **22**, 1368–1373 (1994).
7. Zhao, N., Hashida, H., Takahashi, N., Misumi, Y. & Sakaki, Y. High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression. *Gene* **156**, 207–213 (1995).
8. Nguyen, C. *et al.* Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* **29**, 207–216 (1995).
9. Fodor, S. P. A. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991).
10. Fodor, S. P. *et al.* Multiplexed biochemical assays with biological chips. *Nature* **364**, 555–556 (1993).
11. Pease, A. C. *et al.* Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl Acad. Sci. USA* **91**, 5022–5026 (1994).
12. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
13. Shalon, D., Smith, S. J. & Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).
14. DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
15. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
16. Bowtell, D. D. Options available—from start to finish—for obtaining expression data by microarray. *Nature Genet.* **21**, 25–32 (1999).
17. Edman, C. F. *et al.* Electric field directed nucleic acid hybridization on microchips. *Nucleic Acids Res.* **25**, 4907–4914 (1997).
18. Sosnowski, R. G., Tu, E., Butler, W. F., O'Connell, J. P. & Heller, M. J. Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control. *Proc. Natl Acad. Sci. USA* **94**, 1119–1123 (1997).
19. Gray, D. E., Case-Green, S. C., Fell, T. S., Dobson, P. J. & Southern, E. M. Ellipsometric and interferometric characterization of DNA probes immobilised on a combinatorial array. *Langmuir* **13**, 2833–2842 (1997).
20. Walt, D. R. Bead-based fiber-optic arrays. *Science* **287**, 451 (2000).
21. Michael, K. L., Taylor, L. C., Schultz, S. L. & Walt, D. R. Randomly ordered addressable high-density optical sensor arrays. *Anal. Chem.* **70**, 1242–1248 (1998).
22. Ferguson, J. A., Boles, T. C., Adams, C. P. & Walt, D. R. A fiber-optic DNA biosensor microarray for the analysis of gene expression. *Nature Biotechnol.* **14**, 1681–1684 (1996).
23. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
24. Cho, R. J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
25. The Chipping forecast. *Nature Genet.* **21** (Suppl.), 1–60 (1999).
26. Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H. & Lockhart, D. J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* **15**, 1359–1367 (1997).
27. White, K. P., Rifkin, S. A., Hurban, P. & Hogness, D. S. Microarray analysis of *Drosophila* development during metamorphosis. *Science* **286**, 2179–2184 (1999).
28. Chambers, J. *et al.* DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *J. Virol.* **73**, 5757–5766 (1999).
29. Gingeras, T. R. *et al.* Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic mycobacterium DNA arrays. *Genome Res.* **8**, 435–448 (1998).
30. Lockhart, D. J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
31. Liang, P. & Pardee, A. B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967–971 (1992).
32. Shimkets, R. A. *et al.* Gene expression analysis by transcript profiling coupled to a gene database query. *Nature Biotechnol.* **17**, 798–803 (1999).
33. Ivanova, N. B. & Belyavsky, A. V. Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Res.* **23**, 2954–2958 (1995).
34. Kato, K. Description of the entire mRNA population by a 3′ end cDNA fragment generated by class IIS restriction enzymes. *Nucleic Acids Res.* **23**, 3685–3690 (1995).
35. Bachem, C. W. *et al.* Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant J.* **9**, 745–753 (1996).
36. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
37. Boucherie, H. *et al.* Two-dimensional protein map of *Saccharomyces cerevisiae*: construction of a gene-protein index. *Yeast* **11**, 601–613 (1995).
38. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17**, 994–999 (1999).
39. Mann, M. Quantitative proteomics? *Nature Biotechnol.* **17**, 954–955 (1999).
40. Oda, Y., Huang, K., Cross, F. R., Cowburn, D. & Chait, B. T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl Acad. Sci. USA* **96**, 6591–6596 (1999).
41. Burns, N. *et al.* Large-scale analysis of gene expression, protein localization, and gene disruption in Saccharomyces cerevisiae. *Genes Dev.* **8**, 1087–1105 (1994).
42. Ross-Macdonald, P., Sheehan, A., Roeder, G. S. & Snyder, M. A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **94**, 190–195 (1997).
43. Ross-Macdonald, P. *et al.* Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418 (1999).
44. Niedenthal, R. K., Riles, L., Johnston, M. & Hegemann, J. H. Green fluorescent protein as a marker for gene expression and subcellular localization in budding yeast. *Yeast* **12**, 773–786 (1996).
45. Zong, Q., Schummer, M., Hood, L. & Morris, D. R. Messenger RNA translation state: the second dimension of high-throughput expression screening. *Proc. Natl Acad. Sci. USA* **96**, 10632–10636 (1999).
46. Johannes, G., Carter, M. S., Eisen, M. B., Brown, P. O. & Sarnow, P. Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proc. Natl Acad. Sci. USA* **96**, 13118–13123 (1999).
47. Diehn, M., Eisen, M. B., Botstein, D. & Brown, P. O. Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nature Genet.* **25**, 58–62 (2000).
48. Weinstein, J. N. Fishing expeditions. *Science* **282**, 628–629 (1998).
49. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
50. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression

profiling. *Nature* **403**, 503–510 (2000).
51. Mack, D. H. *et al.* in *Deciphering Molecular Circuitry Using High-Density DNA Arrays* (eds Hihich, E. & Croce, E.) 85–108 (Plenum, New York, 1998).
52. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
53. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
54. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24**, 227–235 (2000).
55. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24**, 236–244 (2000).
56. Fambrough, D., McClure, K., Kazlauskas, A. & Lander, E. S. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* **97**, 727–741 (1999).
57. Lee, S. B. *et al.* The Wilms tumor suppressor WT1 encodes a transcriptional activator of amphiregulin. *Cell* **98**, 663–673 (1999).
58. Harkin, D. P. *et al.* Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. *Cell* **97**, 575–586 (1999).
59. Mei, R. *et al.* Genome-wide detection of allelic imbalance using human SNPs and high density DNA arrays. *Genome Res.* (in the press).
60. Pollack, J. R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet.* **23**, 41–46 (1999).
61. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20**, 207–211 (1998).
62. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
63. Wyrick, J. J. *et al.* Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature* **402**, 418–421 (1999).
64. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
65. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
66. Wen, X. *et al.* Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA* **95**, 334–339 (1998).
67. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
68. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
69. Wolfsberg, T. G. *et al.* Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.* **9**, 775–792 (1999).
70. Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4**, 1293–1301 (1998).
71. Gray, N. S. *et al.* Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* **281**, 533–538 (1998).
72. Rosania, G. R. *et al.* Myoseverin: a microtubule binding molecule with novel cellular effects. *Nature Biotechnol.* **18**, 304–308 (2000).
73. Emmert-Buck, M. R. *et al.* Laser capture microdissection. *Science* **274**, 998–1001 (1996).
74. Bonner, R. F. *et al.* Laser capture microdissection: molecular analysis of tissue. *Science* **278**, 1481–1483 (1997).
75. Simone, N. L., Bonner, R. F., Gillespie, J. W., Emmert-Buck, M. R. & Liotta, L. A. Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends Genet.* **14**, 272–276 (1998).
76. Wang, A. M., Doyle, M. V. & Mark, D. F. Quantitation of mRNA by the polymerase chain reaction. *Proc. Natl Acad. Sci. USA* **86**, 9717–9721 (1989).
77. Dulac, C. Cloning of genes from single neurons. *Curr. Top. Dev. Biol.* **36**, 245–258 (1998).
78. Jena, P. K., Liu, A. H., Smith, D. S. & Wysocki, L. J. Amplification of genes, single transcripts and cDNA libraries from one cell and direct sequence analysis of amplified products derived from one molecule. *J. Immunol. Methods* **190**, 199–213 (1996).
79. Kwoh, D. Y. *et al.* Transcription-based amplification system and detection of amplified human immunodeficiency virus type 1 with a bead-based sandwich hybridization format. *Proc. Natl Acad. Sci. USA* **86**, 1173–1177 (1989).
80. Guatelli, J. C. *et al.* Isothermal, in vitro amplification of nucleic acids by a multienzyme reaction modeled after retroviral replication. *Proc. Natl Acad. Sci. USA* **87**, 7797 (1990).
81. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc. Natl Acad. Sci. USA* **89**, 3010–3014 (1992).
82. Luo, L. *et al.* Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nature Med.* **5**, 117–122 (1999).
83. Cho, R. J. *et al.* Parallel analysis of genetic selections using whole genome oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 3752–3757 (1998).
84. Bulyk, M. L., Gentalen, E., Lockhart, D. J. & Church, G. M. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nature Biotechnol.* **17**, 573–577 (1999).
85. Brent, R. Genomic biology. *Cell* **100**, 169–183 (2000).
86. McAdams, H. H. & Shapiro, L. Circuit simulation of genetic networks. *Science* **269**, 650–656 (1995).
87. McAdams, H. H. & Arkin, A. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* **15**, 65–69 (1999).
88. Bhalla, U. S. & Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science* **283**, 381–387 (1999).
89. Weng, G., Bhalla, U. S. & Iyengar, R. Complexity in biological signaling systems. *Science* **284**, 92–96 (1999).
90. Arkin, A., Ross, J. & McAdams, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**, 1633–1648 (1998).
91. Winzeler, E. *et al.* Functional characterization of the *Saccharomyces cerevisiae* genome by precise deletion and parallel analysis. *Science* **285**, 901–906 (1999).
92. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245–246 (1989).
93. Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
94. Brown, P. O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21**, 33–37 (1999).
95. Ly, D., Lockhart, D. J., Lerner, R. & Schultz, P. G. Mitotic misregulation and human aging. *Science* **287**, 2486–22492 (2000).
96. Cherry, J. M. *et al.* Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* **387**, 67–73 (1997).
97. Ball, C. A. *et al.* Integrating functional genomic information into the *Saccharomyces* Genome Database. *Nucleic Acids Res.* **28**, 77–80 (2000).
98. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
99. Walsh, S., Anderson, M. & Cartinhour, S. W. ACEDB: a database for genome information. *Methods Biochem. Anal.* **39**, 299–318 (1998).
100. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
101. The FlyBase Consortium. The FlyBase Database of the Drosophila Genome Projects and community literature. *Nucleic Acids Res.* **27**, 85–88 (1999).
102. Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* **27**, 50–53 (1999).
103. Iyer, V. & Struhl, K. Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **93**, 5208–5212 (1996).
104. Lee, C. K., Klopp, R. G., Weindruch, R. & Prolla, T. A. Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**, 1390–1393 (1999).
105. Fan, J.-B. *et al.* Parallel genotyping of human SNPs using generic oligonucleotide tag arrays. *Genome Res.* (in the press).
106. Lashkari, D. A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA* **94**, 13057–13062 (1997).
107. Winzeler, E., Lee, B., McCusker, J. & Davis, R. Whole genome genetic typing using high-density oligonucleotide arrays. *Parasitology* **118**, S73–S80 (1999).
108. Winzeler, E. A. *et al.* Direct allelic variation scanning of the yeast genome. *Science* **281**, 1194–1197 (1998).
109. Troesch, A. *et al.* Mycobacterium species identification and rifampin resistance testing with high-density DNA probe arrays. *J. Clin. Microbiol.* **37**, 49–55 (1999).

Acknowledgements

We thank S. Fodor, M. Chee, R. Davis, L. Stryer, E. Lander, H. Dong, L. Wodicka, R. Cho, D. Giang, P. Zarrinkar, C. Barlow, J. Gentry, P. Schultz and R. Abagyan for their on-going help and patience, and B. Geierstanger, G. Hampton and S. Kay for helpful comments and a critical reading of the manuscript.